US006484143B1

(12) **United States Patent**
Swildens et al.

(10) **Patent No.: US 6,484,143 B1**
(45) **Date of Patent: Nov. 19, 2002**

(54) **USER DEVICE AND SYSTEM FOR TRAFFIC MANAGEMENT AND CONTENT DISTRIBUTION OVER A WORLD WIDE AREA NETWORK**

(75) Inventors: **Eric Sven-Johan Swildens**, Mountain View, CA (US); **Richard David Day**, Mountain View, CA (US); **Ajit K. Gupta**, Fremont, CA (US)

(73) Assignee: **Speedera Networks, Inc.**, Santa Clara, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 19 days.

(21) Appl. No.: **09/641,746**

(22) Filed: **Aug. 18, 2000**

Related U.S. Application Data

(60) Provisional application No. 60/166,906, filed on Nov. 22, 1999.

(51) **Int. Cl.**[7] **G06F 17/60**

(52) **U.S. Cl.** **705/1**; 709/223; 709/224; 709/235; 713/213

(58) **Field of Search** 709/223, 224, 709/235, 203, 201; 713/213; 705/1

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,568,471	A *	10/1996	Hershey et al.	370/245
5,742,587	A *	4/1998	Zornig et al.	370/235
6,047,326	A *	4/2000	Kilkki	709/228
6,078,960	A *	6/2000	Ballard	709/203
6,119,143	A *	9/2000	Dias et al.	709/105
6,173,407	B1 *	1/2001	Yoon et al.	713/201
6,185,601	B1 *	2/2001	Wolff	709/105
6,308,148	B1 *	10/2001	Bruins et al.	703/27
6,308,209	B1 *	10/2001	Lecheler	709/224
6,310,881	B1 *	10/2001	Zikan et al.	370/236
6,327,622	B1 *	12/2001	Jindal et al.	709/105
6,337,862	B1 *	1/2002	O'Callaghan et al.	370/392

FOREIGN PATENT DOCUMENTS

WO	WO 02/23309 A2 *	3/2002

OTHER PUBLICATIONS

Drew Jeff, Erhlick Abraham, Is cache King? (internet/web/online service information) Jun. 2001, Business Communications Review, vol. 31, No 6, p. 48.*

Business Editors/Hich-tech Writers, Marimba Launches New Managed service Provider—MSP—Division; Marimba.net Offers Managed Internet Services for ASP's Appliance Vendors, Portals, and Other E-Companies, Oct. 18, 2000, Business Wire, p. 0383.*

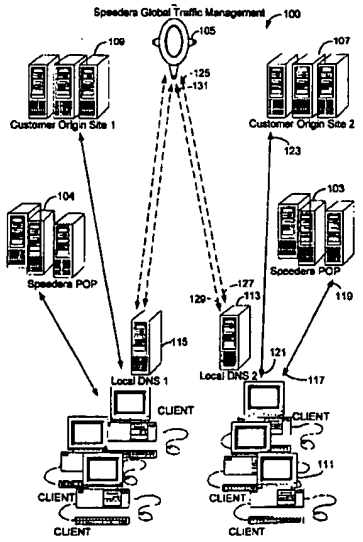
* cited by examiner

Primary Examiner—James P. Trammell
Assistant Examiner—Firmin Backer
(74) *Attorney, Agent, or Firm*—Michael A. Glenn; Kirk D. Wong

(57) **ABSTRACT**

A user interface device and system for providing a shared GTM and CDN (collectively Universal Distribution Network) for a service fee, where the customer or user does not need to purchase significant hardware and/or software features. The present interface device and system allows a customer to scale up its Web site, without a need for expensive and difficult to use hardware and/or software. In a preferred embodiment, the customer merely pays for a service fee, which can be fixed, variable, lump some, or based upon a subscription model using the present system. The present device and system are preferably implemented on a system including a novel combination of global traffic management and content distribution.

22 Claims, 16 Drawing Sheets



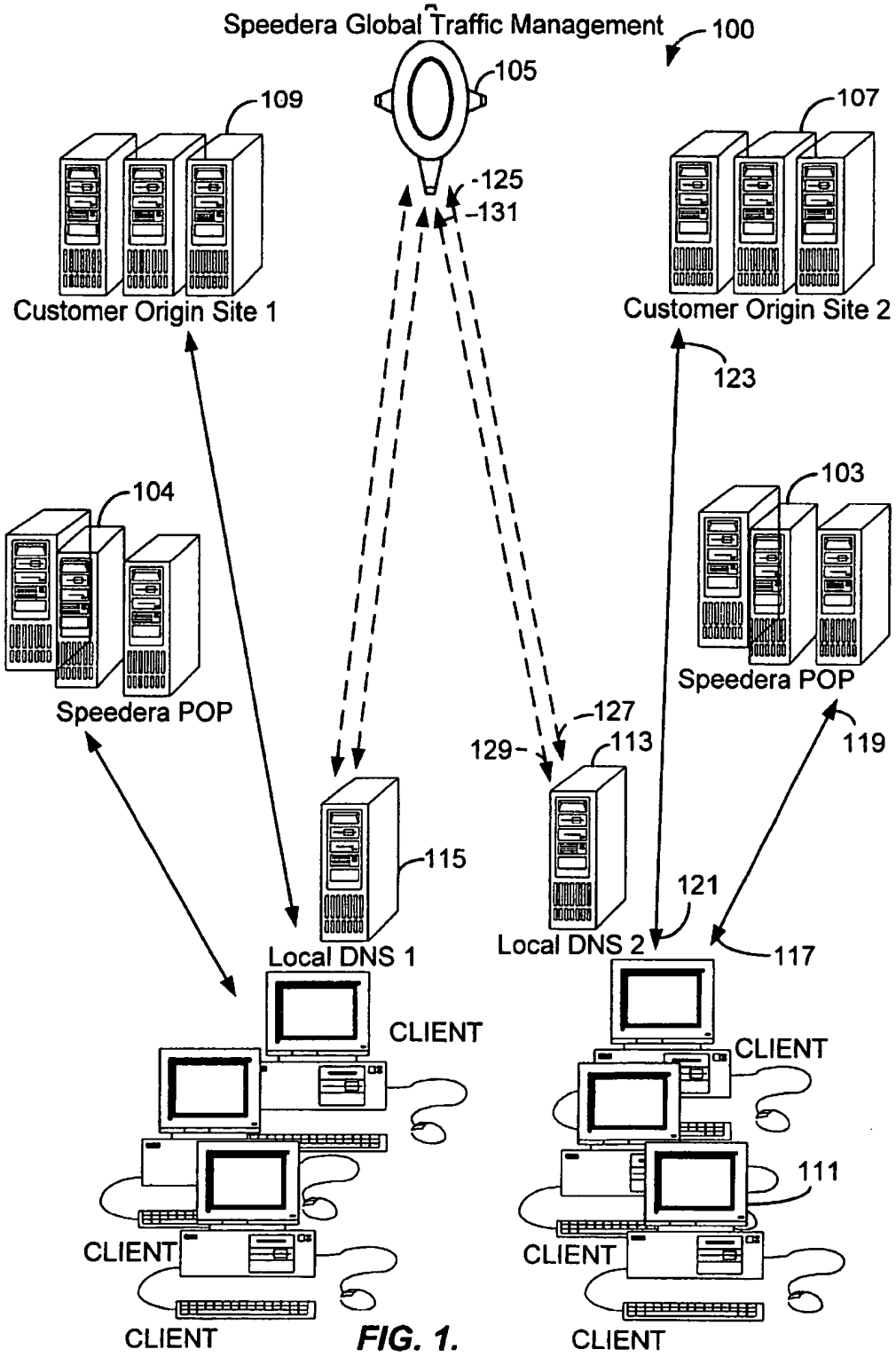

Speedera Global Traffic Management
100
105
109
107
125
131
Customer Origin Site 1
Customer Origin Site 2
123
104
103
Speedera POP
Speedera POP
127
129
113
119
115
121
Local DNS 1
Local DNS 2
117
CLIENT
CLIENT
111
CLIENT
CLIENT
CLIENT

FIG. 1.

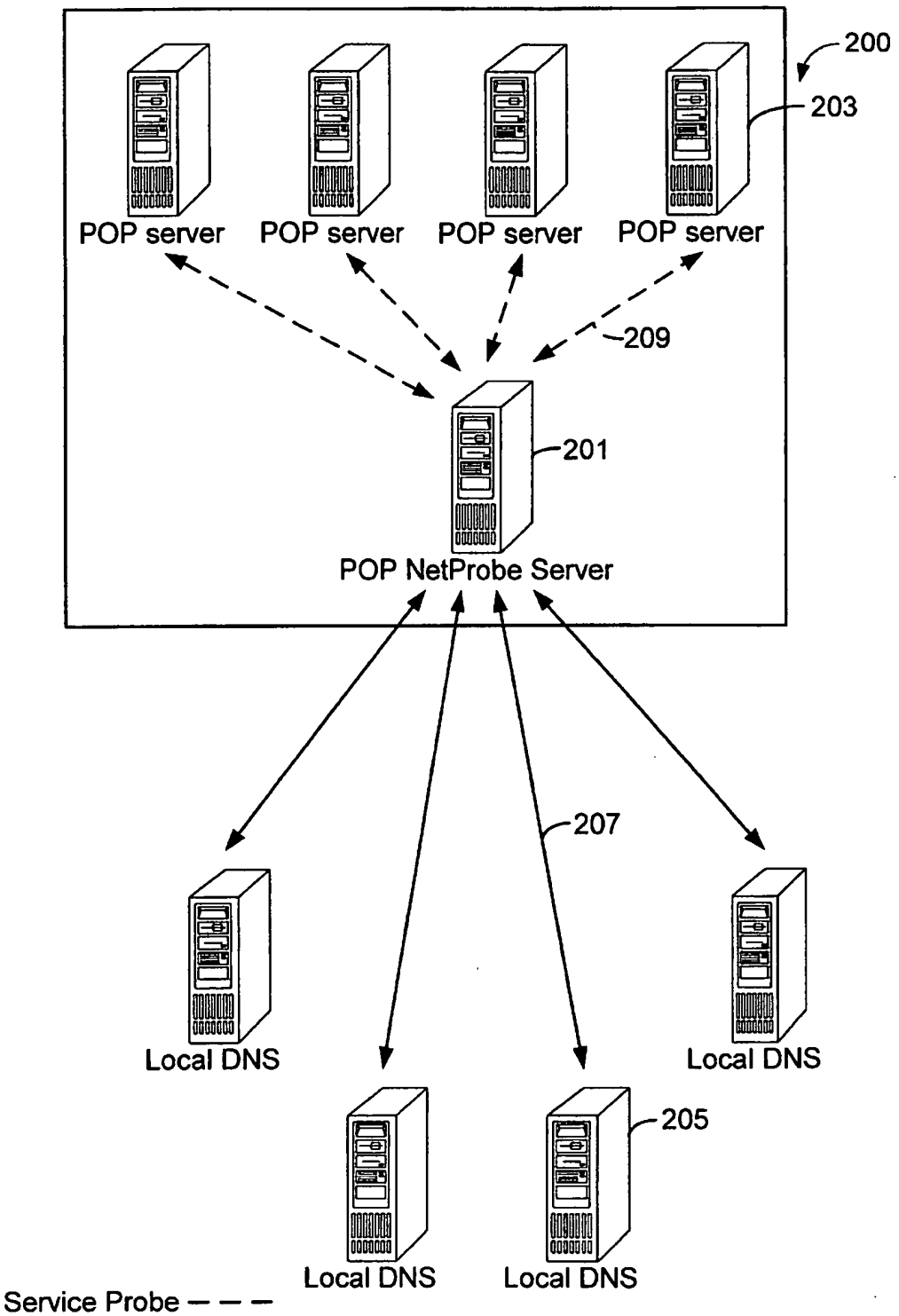

200
203
POP server
POP server
POP server
POP server
209
201
POP NetProbe Server
207
Local DNS
Local DNS
Local DNS
Local DNS
205
Service Probe
Latency Probe

FIG. 2.

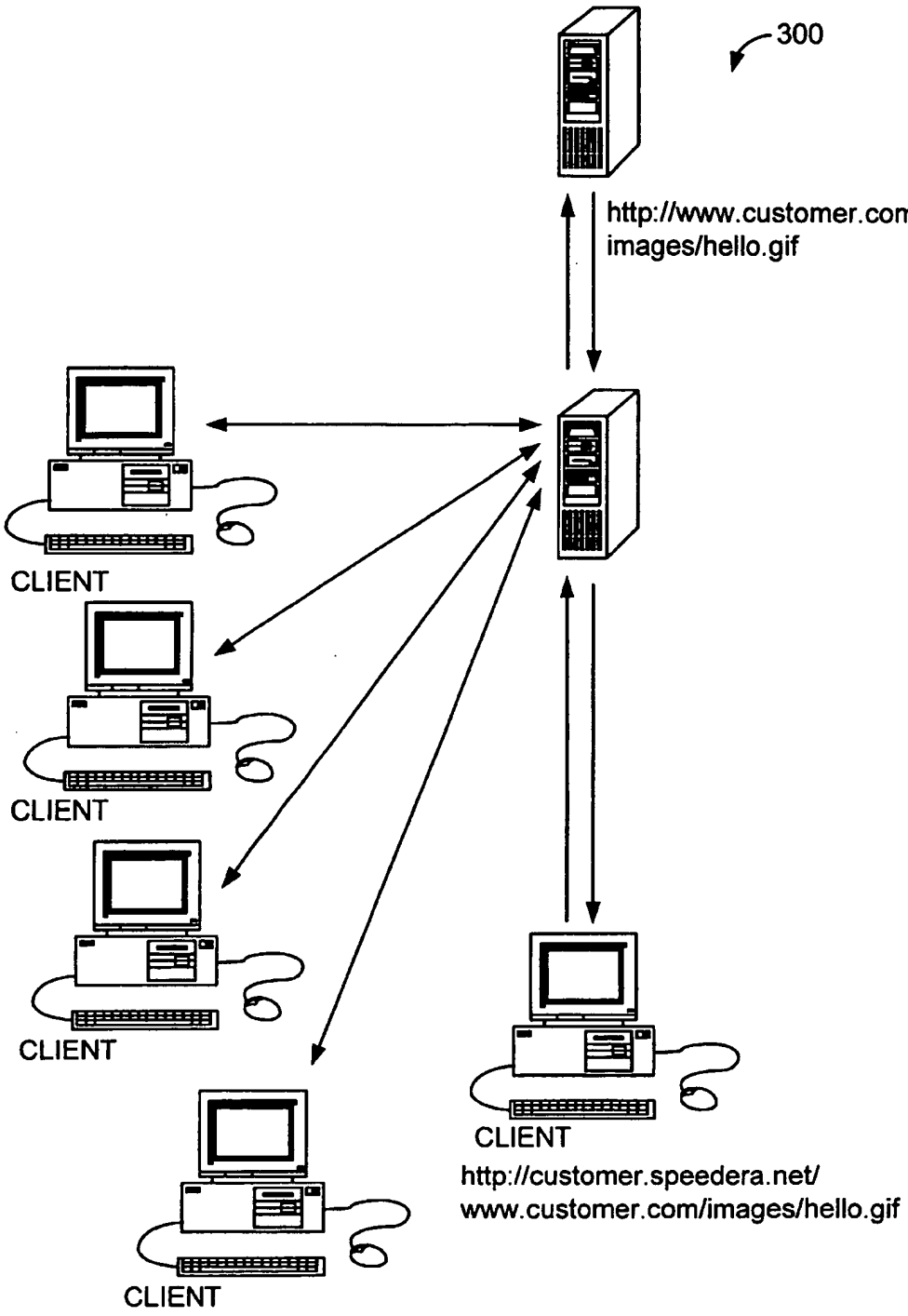

300
http://www.customer.com/
images/hello.gif
CLIENT
CLIENT
CLIENT
CLIENT
CLIENT
http://customer.speedera.net/
www.customer.com/images/hello.gif

FIG. 3.

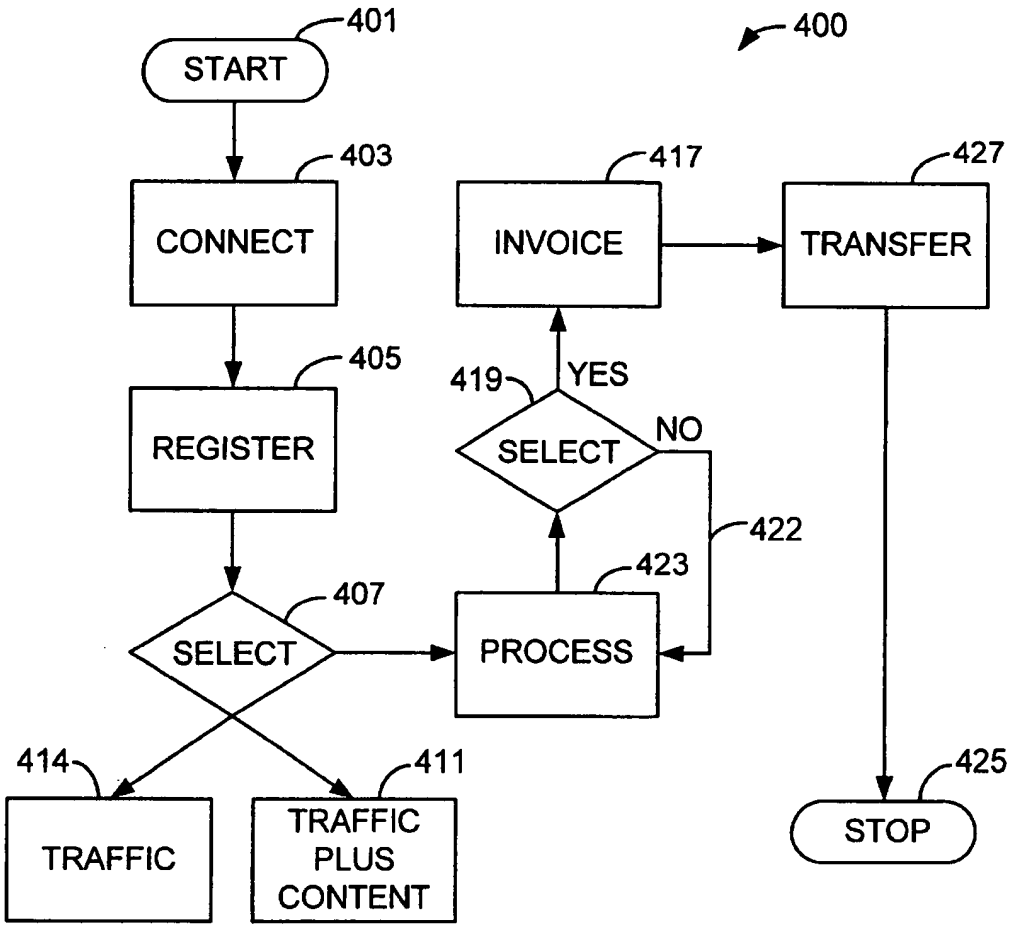

400
401
START
403
CONNECT
405
REGISTER
407
SELECT
414
TRAFFIC
411
TRAFFIC PLUS CONTENT
417
INVOICE
419
YES
SELECT
NO
422
423
PROCESS
427
TRANSFER
425
STOP

FIG. 4.

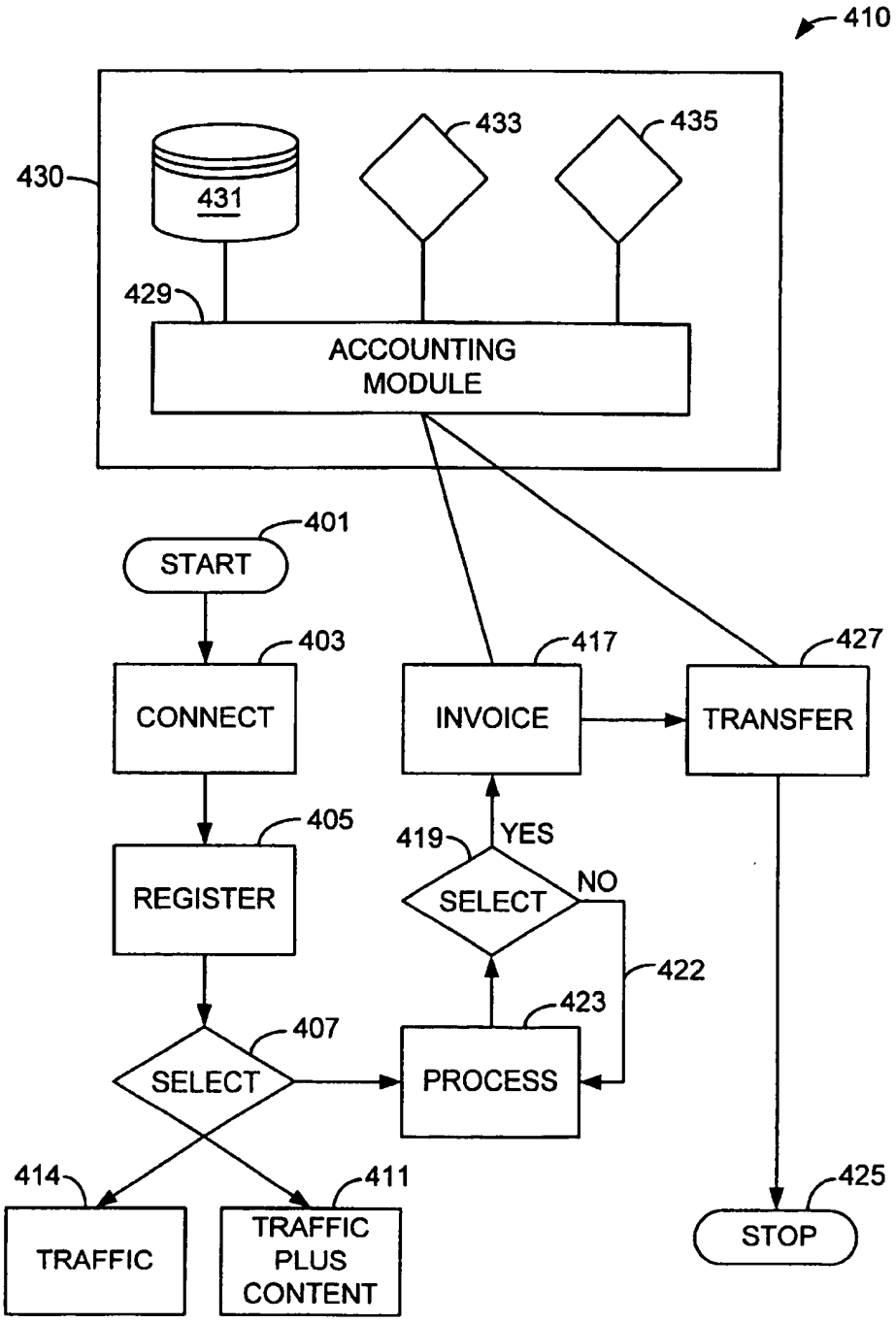

410
430
431
433
435
429
ACCOUNTING MODULE
401
START
403
CONNECT
405
REGISTER
407
SELECT
414
TRAFFIC
411
TRAFFIC PLUS CONTENT
417
INVOICE
419
YES
SELECT
NO
422
423
PROCESS
427
TRANSFER
425
STOP

FIG. 4A.

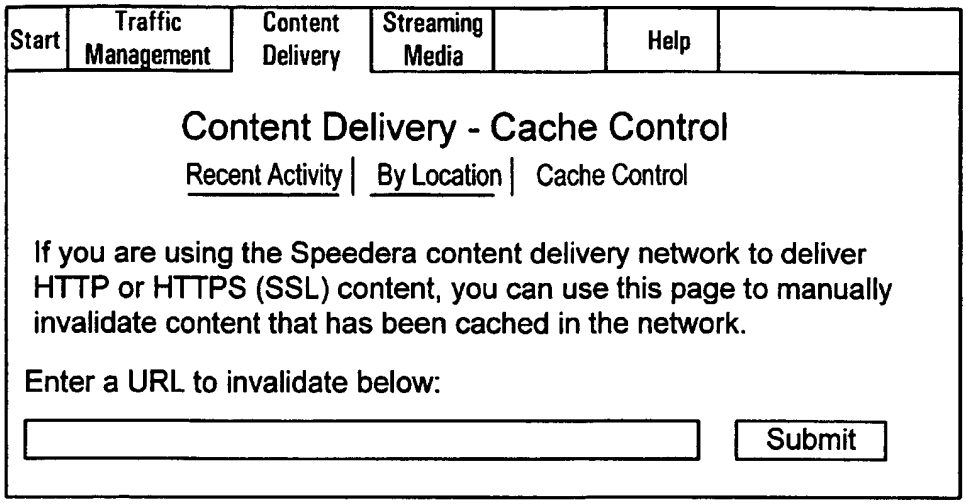

Start
Traffic Management
Content Delivery
Streaming Media
Help
Content Delivery - Cache Control
Recent Activity | By Location | Cache Control
If you are using the Speedera content delivery network to deliver HTTP or HTTPS (SSL) content, you can use this page to manually invalidate content that has been cached in the network.
Enter a URL to invalidate below:
Submit

FIG. 5A.

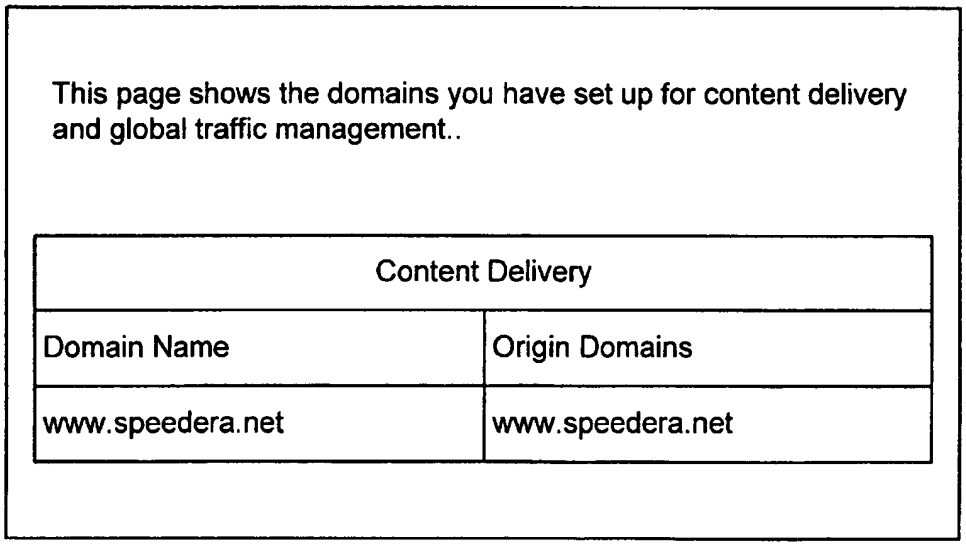

This page shows the domains you have set up for content delivery and global traffic management..
Content Delivery
Domain Name
Origin Domains
www.speedera.net
www.speedera.net

FIG. 5B.

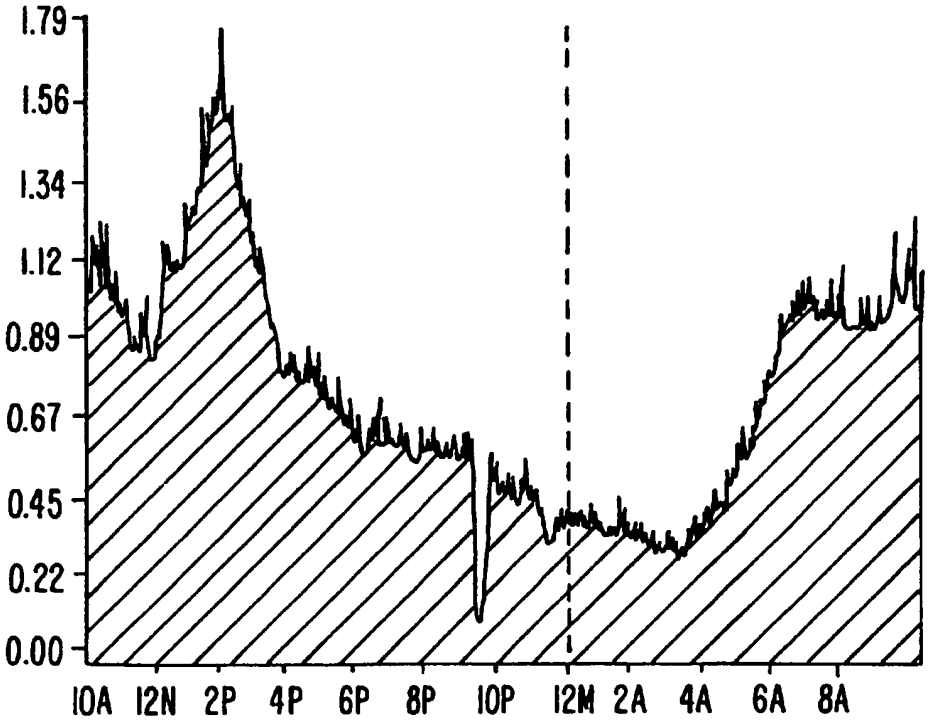

1.79
1.56
1.34
1.12
0.89
0.67
0.45
0.22
0.00
10A 12N 2P 4P 6P 8P 10P 12M 2A 4A 6A 8A

FIG. 5C.

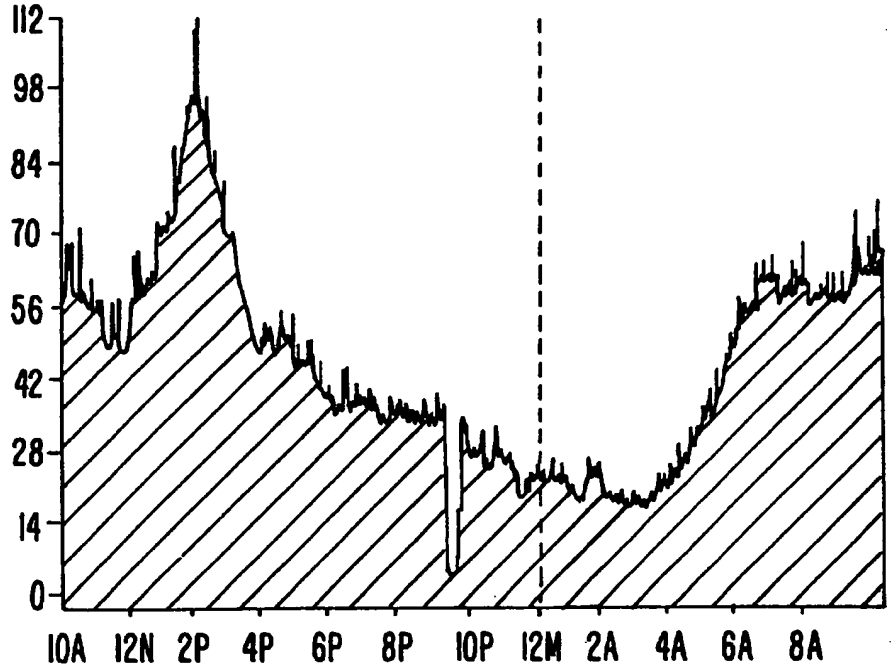


FIG. 5D.

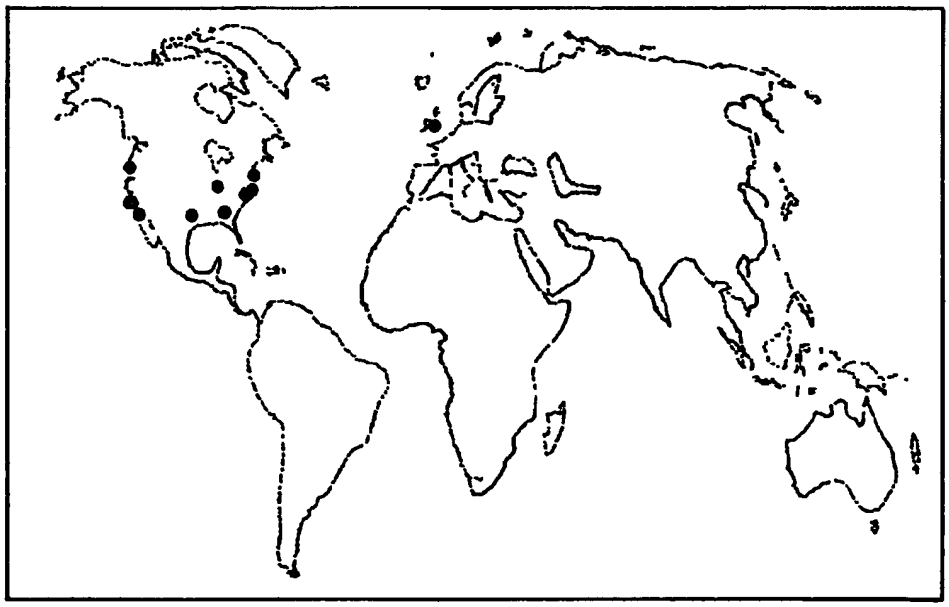


FIG. 5E.

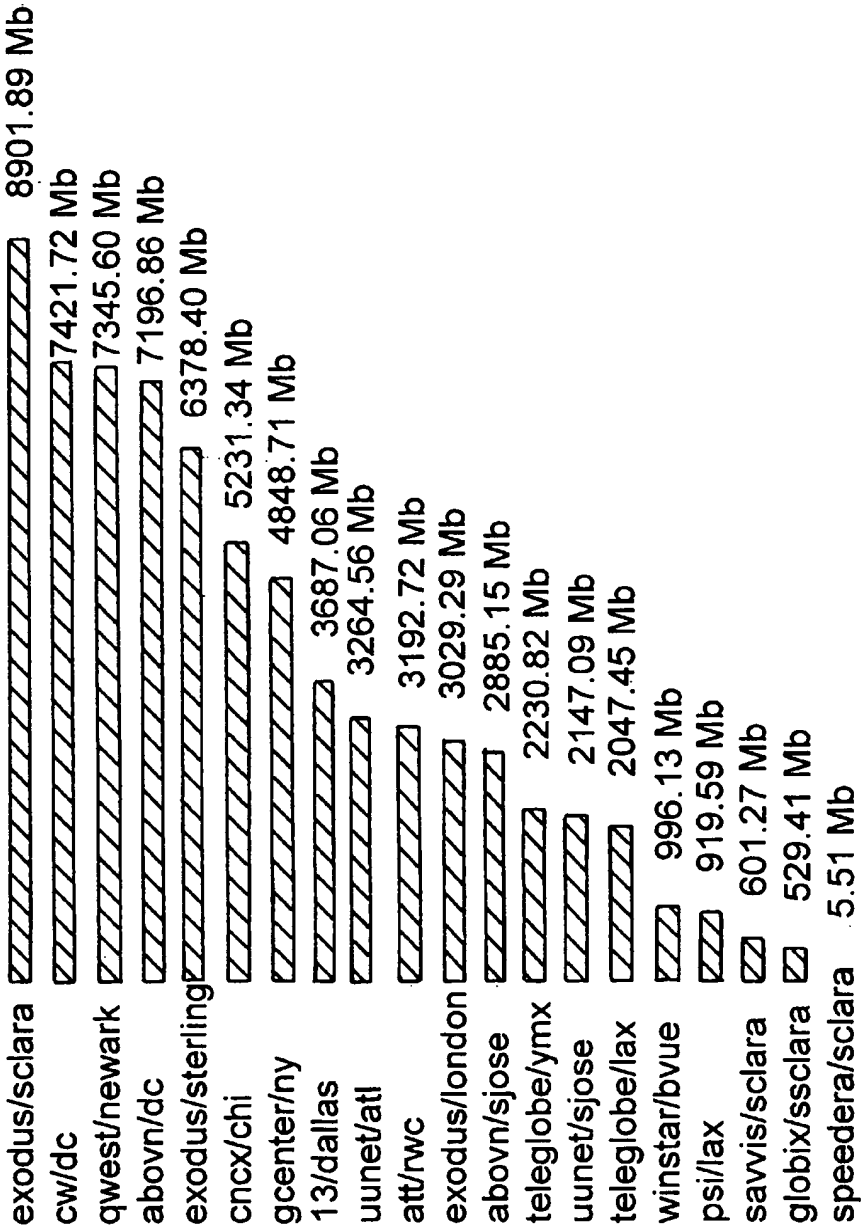

exodus/sclara 8901.89 Mb
cw/dc 7421.72 Mb
qwest/newark 7345.60 Mb
abovn/dc 7196.86 Mb
exodus/sterling 6378.40 Mb
cncx/chi 5231.34 Mb
gcenter/ny 4848.71 Mb
13/dallas 3687.06 Mb
uunet/atl 3264.56 Mb
att/rwc 3192.72 Mb
exodus/london 3029.29 Mb
abovn/sjose 2885.15 Mb
teleglobe/ymx 2230.82 Mb
uunet/sjose 2147.09 Mb
teleglobe/lax 2047.45 Mb
winstar/bvue 996.13 Mb
psi/lax 919.59 Mb
savvis/sclara 601.27 Mb
globix/ssclara 529.41 Mb
speedera/sclara 5.51 Mb

FIG. 5F.

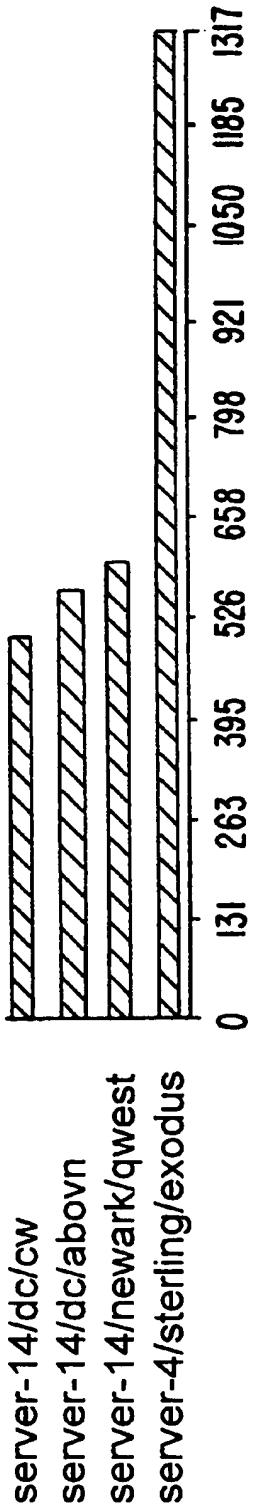
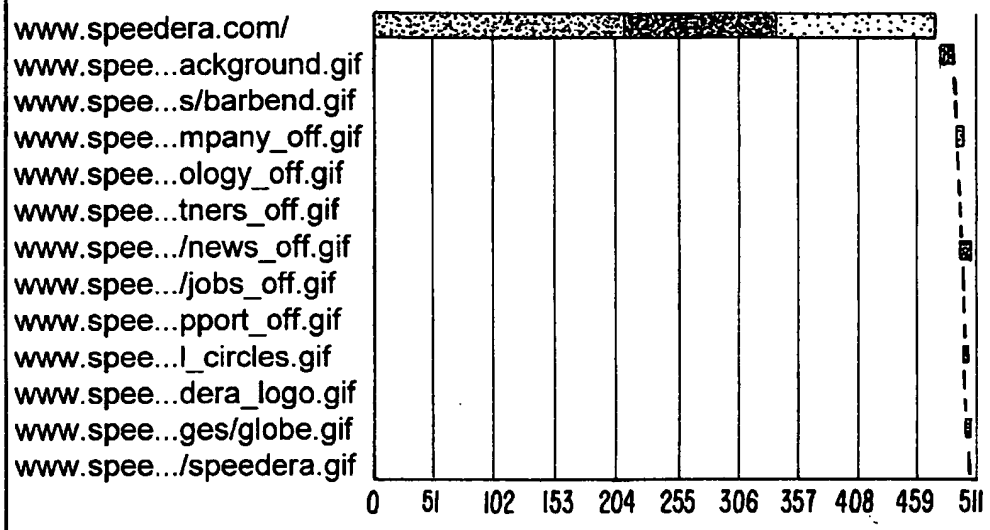

server-14/dc/cw
server-14/dc/abovn
server-14/newark/qwest
server-4/sterling/exodus
0
131
263
395
526
658
798
921
1050
1185
1317

FIG. 5G.

Location server: 14 dc, cw (204./1.35.146)



URL

IP Address	ERR	HRC	LEN	CHK	STT	DRT	COT	DST	FNT	END
www.speedera.com/										
209.24.35.130	0	200	7586	632552	0	0	214	130	136	482
www.speedera.net/www.speedera.com/images/background.gif										
204.71.35.135	0	200	87	8934	482	12	0	1	0	496
www.speedera.net/www.speedera.com/images/barbend.gif										
204.71.35.134	0	200	3009	357219	494	1	1	1	0	498

FIG. 5H.

Global Traffic Management		
Domain Name	Location	IP Addresses
customer.speedera.net	UUNET	a,b,c,d
	VSNL/delhi	w,x,y,z

FIG. 6A.

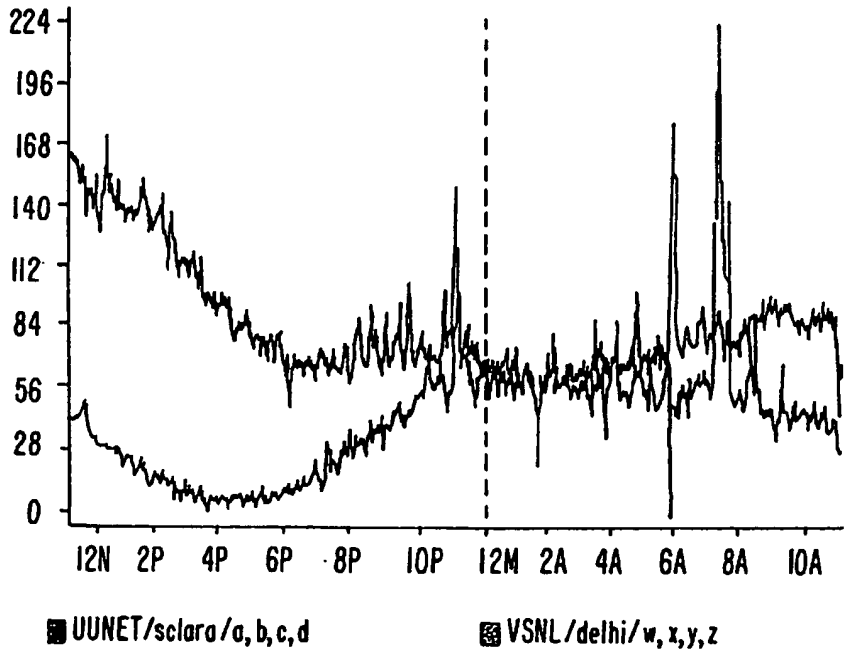


FIG. 6B.

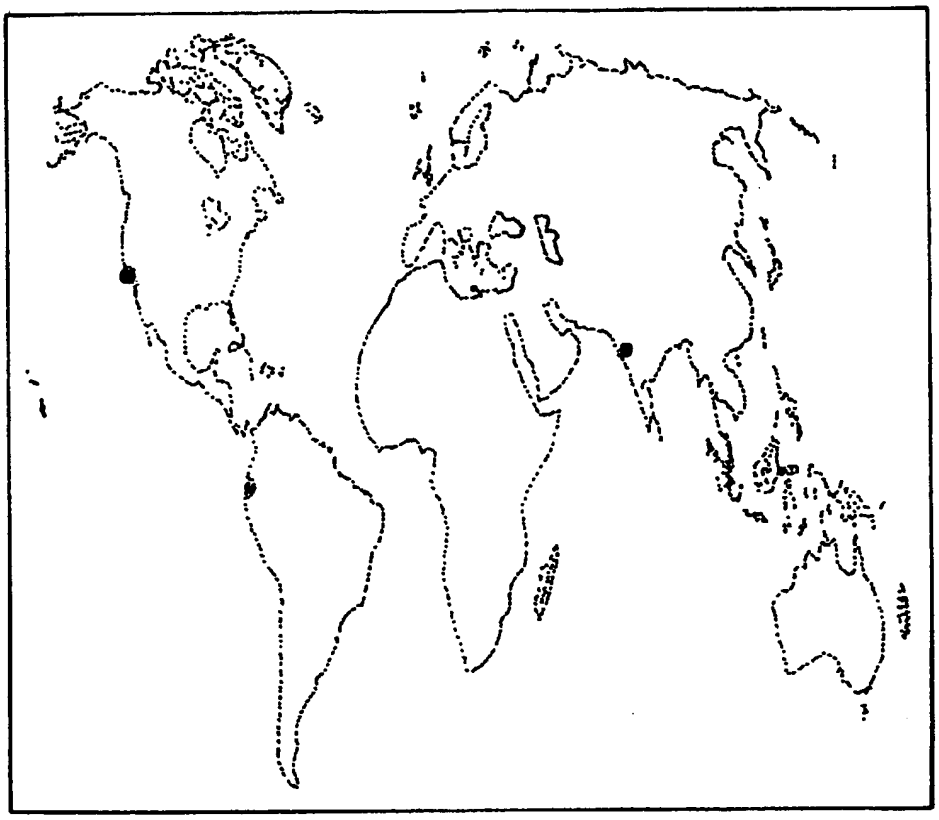

SERVER SCHEDULING BY LOCATION - LAST 24 HOURS
UUNET/sclara
131374
VSNL/delhi
66746

FIG. 6C.

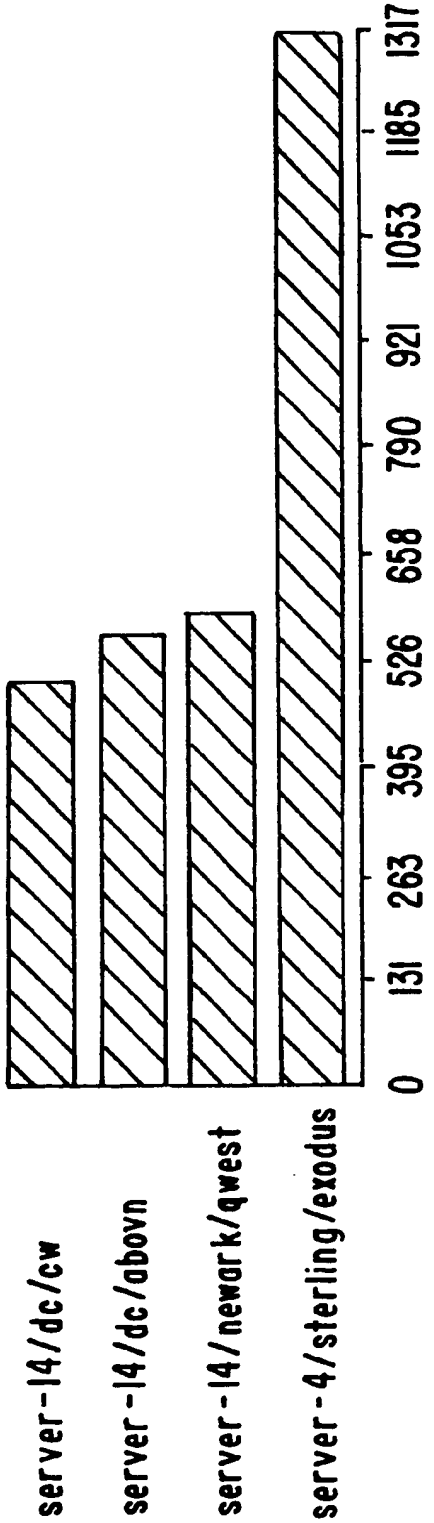
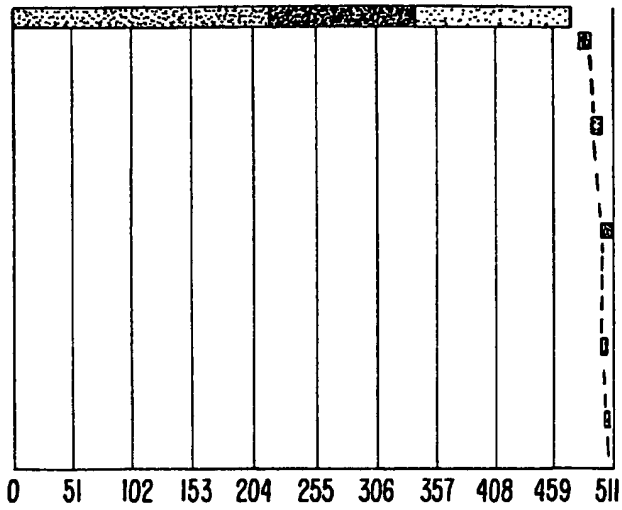

0
131
263
395
526
658
790
921
1053
1185
1317
server-14/dc/cw
server-14/dc/abovn
server-14/newark/qwest
server-4/sterling/exodus

FIG. 6D.

Location server: 14 dc, cw (204./1.35.146)

www.speedera.com/
www.spee...ackground.gif
www.spee...s/barbend.gif
www.spee...mpany_off.gif
www.spee...ology_off.gif
www.spee...tners_off.gif
www.spee.../news_off.gif
www.spee.../jobs_off.gif
www.spee...pport_off.gif
www.spee...l_circles.gif
www.spee...dera_logo.gif
www.spee...ges/globe.gif
www.spee.../speedera.gif



URL

IP Address	ERR	HRC	LEN	CHK	STT	DRT	COT	DST	FNT	END
www.speedera.com/										
209.24.35.130	0	200	7586	632552	0	0	214	130	136	482
www.speedera.net/www.speedera.com/images/background.gif										
204.71.35.135	0	200	87	8934	482	12	0	1	0	496
www.speedera.net/www.speedera.com/images/barbend,gif										
204.71.35.134	0	200	3009	357219	494	1	1	1	0	498

FIG. 6E.

USER DEVICE AND SYSTEM FOR TRAFFIC MANAGEMENT AND CONTENT DISTRIBUTION OVER A WORLD WIDE AREA NETWORK

CROSS-REFERENCES TO RELATED APPLICATIONS

The present application claims priority to U.S. Provisional Application No. 60/166,906 filed Nov. 22, 1999, commonly owned, and hereby incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

The present invention relates to world wide area networking. More particularly, the invention provides a technique including a user interface device and system for using a global traffic management system coupled to a plurality of content servers for a service fee. But it would be recognized that the invention has a much broader range of applicability. For example, the invention can also be applied to other types of networks, and the like.

The Internet is a world wide "super-network" which connects together millions of individual computer networks and computers. The Internet is generally not a single entity. It is an extremely diffuse and complex system over where no single entity has complete authority or control. Although the Internet is widely know for one of its ways of presenting information through the World Wide Web (herein "Web"), there are many other services currently available based upon the general Internet protocols and infrastructure.

The Web is often easy to use for people inexperienced with computers. Information on the Web often is presented on "pages" of graphics and text that contain "links" to other pages either within the same set of data files (i.e., Web site) or within data files located on other computer networks. Users often access information on the Web using a "browser" program such as one made by Netscape Communications Corporation (now America Online, Inc.) of Mountain View, Calif. or Explorer™ from Microsoft Corporation of Redmond, Wash. Browser programs can process information from Web sites and display the information using graphics, text, sound, and animation. Accordingly, the Web has become a popular medium for advertising goods and services directly to consumers.

As time progressed, usage of the Internet has exploded. There are literally millions of users on the Internet. Usage of the Internet is increasing daily and will eventually be in the billions of users. As usage increases so does traffic on the Internet. Traffic generally refers to the transfer of information from a Web site at a server computer to a user at a client computer. The traffic generally travels through the world wide network of computers using a packetized communication protocol, such as TCP/IP. Tiny packets of information travel from the server computer through the network to the client computer. Like automobiles during "rush hour" on Highway 101 in Silicon Valley, the tiny packets of information traveling through the Internet become congested. Here, traffic jams which cause a delay in the information from the server to the client occur during high usage hours on the Internet. These traffic jams lead to long wait times at the client location. Here, a user of the client computer may wait for a long time for a graphical object to load onto his/her computer.

From the above, it is seen that an improved way to transfer information over a network is highly desirable.

SUMMARY OF THE INVENTION

According to the present invention, a technique including a user interface device and system for global traffic management and content distribution is provided. In an exemplary embodiment, the method is applied to a world wide network of computers, such as the Internet or an internet.

In a specific embodiment, the invention provides a service based system for traffic management and content distribution for a plurality of users over a world wide network of computers. The system includes a global traffic management device coupled to a world wide area network. The global traffic management device being provided to load balance across multiple origin sites. The system also has a content delivery network coupled to the global traffic management device. The content delivery network provides support content distribution and delivery of streaming media. The system also has a computing device including a computer memory coupled to the global traffic management device. The system also has an accounting module coupled to the computing device. The accounting module tracks a usage of the global traffic management device and the content delivery network for a customer of the global traffic management device and the content delivery network to determine a service fee for the usage based upon a period time frequency.

Many benefits are achieved by way of the present invention over conventional techniques. For example, the present invention can be implemented using conventional hardware and software in an easy manner in some embodiments. The invention can also be applied to conventional Web hosting sites. In other aspects, the invention provides an easy way for a Web site to using a content distribution network without spending capital costs. Depending upon the embodiment, one or more of these benefits may be achieved. These and other benefits will be described in more throughout the present specification and more particularly below.

Various additional objects, features and advantages of the present invention can be more fully appreciated with reference to the detailed description and accompanying drawings that follow.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a simplified diagram of a system according to an embodiment of the present invention;

FIG. **2** is a more detailed diagram of probes used in the system according to an embodiment of the present invention;

FIG. **3** is a more detailed diagram of a caching sequence used in the system according to an embodiment of the present invention;

FIG. **4** is a simplified flow diagrams of methods according to embodiments of the present invention;

FIG. **4A** is a simplified system diagram according to an embodiment of the resent invention;

FIGS. **5A** to **5H** are simplified diagrams of content delivery network ccording to an embodiment of the present invention; and

FIGS. **6A** to **6E** are simplified diagrams of global traffic management system according to an embodiment of the present invention

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

According to the present invention, a technique including a user interface device and system for global traffic management and content distribution is provided. In an exemplary embodiment, the method is applied to a world wide network of computers, such as the Internet or an internet.

In a specific embodiment, the invention provides a user interface device and system for providing a shared GTM and

CDN (collectively Universal Distribution Network) for a service fee, where the customer or user does not need to purchase significant hardware and/or software features. The present interface device and system allows a customer to scale up its Web site, without a need for expensive and difficult to use hardware and/or software. In a preferred embodiment, the customer merely pays for a service fee, which can be fixed, variable, lump some, or based upon a subscription model using the present system. The present device and system are preferably implemented on a system including a novel combination of global traffic management and content distribution.

An overall system diagram **100** is illustrated in FIG. 1. The diagram is merely an example, which should not unduly limit the scope of the claims herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. As shown, the system **100** includes a variety of features to defined the Universal Delivery Network (UDN). The UDN has a combined content delivery network **103** and **104** and a global traffic management network **105**, which are coupled to each other. This eliminates the need for independent CDN and GTM solutions. The UDN can be implemented as a single outsourced solution or service to a customer. When deployed across the WAN, it creates a unified network that provides a universal solution for content routing and high availability delivery.

Customers can leverage the size, scope, and location of the UDN to store content such as HTML, images, video, sound and software for fast and highly available access by clients. The network can also incorporate customer origin sites **107, 109** that will then benefit from shared load balancing and traffic management. Customers with generated content, such as search engines, auctions and shopping carts, can use the latter feature to add their own content servers to the network. In some embodiments, the system typically requires no software or hardware to be installed or run at a customer site. A web interface is available for display of the network's current status as well as historical statistics on a per customer basis.

The system functions by mapping hostnames, such as www.customer.com to a customers origin servers **107** and **109**. The local DNS **113** queries the traffic management system **105** for name resolution of the customers web site and receives a response specifying the server best suited to handle the request, either customer origin servers **107** or servers **103** located in the UDN. When the client **111** requests a customer homepage, tags within the HTML direct the imbedded static content to the network of cache servers **103** and **104**. In this example the static content may be tagged with a domain name like customer.speedera.com. Each local DNS in the example is directed to a different resource for each hostname based on several factors, such as proximity to the resource, network congestion, and server load.

In this example, www.customer.com is mapped to the customer origin servers represented by customer origin Sites **1 109** and **2 107**. Customer.speedera.net is mapped to a collection of delivery nodes represented by point of presence servers, i.e., POPs **103, 104**. As merely an example, a method for using such a UDN is provided below.

1. The client **111** requests a customer home page: www.customer.com from a local DNS **113**.
2. The local DNS **113** queries the traffic management system **105** for name and address resolution and receives a reply **125, 127** indicating the optimal customer origin site to retrieve the homepage **131**. In this step, the traffic management system still looks at many if not all factors; network health, server health, packet loss, cost, etc. to determine the optimal customer origin site.
3. The client connects to the site and retrieves the home page (solid blue line) **123, 121**.
4. An object with the image tag specifying http://customer.speedera.net/www.customer.com/hello.gif is found in the HTML of the homepage.
5. The local DNS queries the traffic management system for name and address resolution.
6. The traffic management system looks **129, 131** at factors such as network performance and server load and returns the address of the POP best suited to serve the requested content.
7. The client then retrieves the content from the specified delivery node **117, 119**.

This sequence of steps is merely illustrative. The steps can be performed using computer software or hardware or a combination of hardware and software. Any of the above steps can also be separated or be combined, depending upon the embodiment. some cases, the steps can also be changed in order without limiting the scope of the invention claimed herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. Details of each of the features noted above are more fully described below.

The DNS server (DNS) can be thought of as the traffic director of the system. It contains a mapping of where resources (grouped by hostnames) have been allocated as well as the current state of each resource and their availability to each client. It receives the static information (the mappings) from the configuration file and the dynamic information (resource availability) from the probes. The configuration file also instructs the DNS server how to weight the various criteria available when making its decisions. The DNS is a fully functional DNS server and is compatible with current versions of BIND. Decision criteria cover such areas as resource availability, resource load, latency, static mapping configuration, persistence requirements, fail over logic, weighting parameters, and others, each of which can be alone or combined.

Multiple DNS servers are deployed to provided high availability. The DNS servers are spread throughout the network to avoid single points of failure. The DNS server was designed from the beginning with the ability to proxy requests. This proxy ability combined with algorithms to divide client latency and persistence information across a group of DNS servers greatly reduces the problems associated with WAN replication and synchronization. In the event a request arrives at a DNS server that is not authoritative for this client, the DNS can proxy the request to any number of servers to find an authoritative answer.

The DNS server logs both request and operational data to the database for subsequent viewing. Both real-time and historical views are available. The request data allows the administrator and customer to see to the number of requests directed to each POP on a per hostname basis. The operational data provides statistics about the DNS server and would typically only be viewed by the administrator.

The present system also uses one or more probes to detect information about certain criteria from the network. There are probes including a NetProbes, a ServiceProbe and a LatencyProbe. ServiceProbes test local server resources while LatencyProbes conduct network round trip tests to

clients. Each POP in the network is assigned a ServiceProbe and a LatencyProbe—these can be separate machines but in most cases, the same machine will perform both types of probe.

The NetProbes are responsible for providing the traffic management system with service and latency metrics. The metrics are reported to the DNS server and LogServers. FIG. **2** is a simplified diagram **200** of these probes according to embodiments of the present invention. This diagram is merely an example which should not limit the scope of the claims herein. One of ordinary skill in the art would recognize many variations, alternatives, and modifications. The diagram **200** includes a POP **201**, which includes a NetProbes server. Service probes monitor the POP servers to test the availability and load of the services they support. The latency probe tests the round trip time between the POP and the DNS servers.

A ServiceProbe determines service metric information for servers in the UDN and reports them to the DNS server. Service metrics are one of the decision criteria used by the DNS to make its routing determinations. Each server in the UDN supports one or more services—a web server provides HTTP service, a FTP server provides FTP service. The service probe uses various approaches for gathering data—a service test and statistical monitoring. The value of a service metric is dependent on the metric type and it's implementation.

The HTTP service is an example of the service test approach. Rather then try to test the individual characteristics of a server that may have an impact on performance, the service itself is evaluated as a user would experience it, in order to determine its response time and validity. LOADP, a process running on each server, is implemented as a statistical monitor and is used as a generic service for testing purposes. LOADP provides direct measurement of many system parameters including CPU load, memory usage, swap and disk status, and is used in load balancing decisions.

Hostnames in the system are mapped to service types. This allows a given server to support multiple services and be evaluated independently for each of them. When a request for a particular hostname arrives at a DNS, the service associated with that hostname is compared on each of the machines to find the best-suited server. The data from the probes are sent to both the DNS as well as the database. By sending the data to the database, it allows the performance of the network to be viewed in real time as well as over a period of time.

Every server in the UDN is housed in a POP and each POP has a Latency Probe assigned to it, as shown. The Latency Probes determine the latency from their location to other locations on the Internet (specifically to client DNS' requesting name resolution). The DNS' use this information in determining the best-suited server for a particular request. The list of locations that are used in order to determine the latency is driven by the DNS. When it is determined by a DNS server that its current information regarding latency between "x" number of POPs and a client's local DNS has become stale, it will instruct the probe for that particular POP to recalculate the latency.

The probes utilize a collection of methods to determine the latency based on cost. The probe uses the least expensive method first and moves on to more expensive methods if no results are determined. The probe is designed so new methods can be plugged in as they are developed. The methods can be either active or passive and are prioritized based on accuracy. Active methods may take the form of ping or traceroute but are typically more sophisticated. Passive methods could reference local BGP tables to determine cost metrics.

The individual latency data is sent to the DNS servers while operational data of each method, their success rates, etc are sent to the database. This allows the current and new methods to be monitored and managed. LatencyProbes perform latency tests to the local client DNS (LDNS). The LatencyProbes build a table of LDNS' to test over time, receiving the list of which DNS client IP addresses to probe from the DNS Servers in the network.

In a specific embodiment, the delivery nodes are the edge delivery servers of the network. The invention can support any types of IP based delivery servers including but not limited to HTTP, SSL, FTP, Streaming, NNTP, and DNS servers. In preferred embodiments, the invention uses an HTTP server and SSL cache server. The HTTP and SSL servers are identical with the exception of the encryption used on the data to and from the SSL cache in some embodiments. These servers have a proxy component that allows them to fill their cache by making requests to an origin site if a requested object is not in the cache. A method according to the invention can be briefly described as follows in reference to the simplified diagram **300** of FIG. **3**:

1. An initial user makes a request to the cache for an object http://customer.speedera.net/www.cutomer.com/images/test.gif (Step 1);
2. The cache, discovering that it does not have the object, will find the name of the origin site in the URL (www.customer.com) and make a request to the origin site for /images/test.gif (Step 2);
3. When the cache receives the object it is saved on disk and memory and returned to the initial user. Subsequent users who make requests for the same object will be satisfied by the cache directly (Step 3).

This sequence of steps is merely illustrative. The steps can be performed using computer software or hardware or a combination of hardware and software. Any of the above steps can also be separated or be combined, depending upon the embodiment. In some cases, the steps can also be changed in order without limiting the scope of the invention claimed herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives.

Other protocols will work in a similar fashion unless there is a time concern with loading the first request. An example of this is a live streaming event or large file downloads (patches or video on demand). In these cases the caches may be pre-filled with the data that they need to serve. This pre-filling may take place over terrestrial lines or via satellite in some cases. Statistics about data delivered from the delivery nodes are reported through the logging system to the database for subsequent viewing and analysis.

The system also has a user interface. Here, engineering staff as well as customers can login to monitor and administer the network access from nearly any Internet connected web browser (with proper authentication). The user interface includes tables and graphs from the database. Data arrives at the user interface through the Logging System. This system has two parts: Log Distributor daemons and Log Collector daemons. This daemon monitors a defined directory for completed log files. Log files are defined as complete when they reach a defined size or age. A logging API which all resources share controls the definitions of size and age. When the Log Distributor finds completed log files it is able to send them back to one of many Log Collector daemons for insertion in the database.

As noted, the present network has many advantages. The network has as comprehensive, extensible, multi-faceted global traffic management system as its core, which is

coupled to a content delivery network. Further details of the present content delivery network and global traffic management device are provided below. According to the present invention, a method for providing service to customers is provided. Details of such service are provided below.

FIG. **4** is a simplified flow diagram of a novel service method **400** according to an embodiment of the present invention. The diagram is merely an example, which should not unduly limit the scope of the claims herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. As shown, the method begins at start, step **401**. The method connects (step **403**) a client to a server location through a world wide network of computers. The world wide network of computers can include an internet, the Internet, and others. The connection occurs via a common protocol such as TCP/IP. The client location is coupled to a server, which is for a specific user. The user can be any web site or the like that distributes content over the network. As merely an example, the user can be a portal such as Yahoo! Inc. Alternatively, the user can be an electronic commerce site such as Amazon-.com and others. Further, the user can be a health site. Information sites include the U.S. Patent Office web site, educational sites, financial sites, adult entertainment sites, service sites, business to business commerce sites, etc. There are many other types of users that desire to have content distributed in an efficient manner.

In a specific embodiment, the user registers its site on the server, which is coupled to a content distribution server coupled to a global traffic management server. The user registers to select (step **407**) a service from the server. The service can be either a traffic management service (step **414**) or a traffic management service and content distribution service (step **411**). Depending upon the embodiment, the user can select either one and does not need to purchase the capital equipment required for either service. Here, the user merely registers for the service and pays a service fee. The service fee can be based upon a periodic time frequency or other parameter, such as performance, etc. Once the service has been requested, the user performs some of the steps noted herein to use the service.

Next, the method processes (step **423**) the user's request and allows the user to use the content distribution network and/or global traffic management network, where the user's web pages are archives and distributed through the content distribution network in the manner indicated herein. The user's web site should become more efficient from the use of such networks. Once a periodic time frequency or other frequency has lapsed (step **419**), the method goes to an invoicing step, step **417**. The invoicing step sends (step **427**) an invoice to the user. Alternatively, the process continues until the periodic time frequency for the designated service lapses via line **422**. The invoice can be sent via U.S. mail, electronic mail, or the like. The method stops, step **425**. Alternatively, the invoicing step can deduct monetary consideration through an electronic card, e.g., debit card, credit card.

This sequence of steps is merely illustrative. The steps can be performed using computer software or hardware or a combination of hardware and software. Any of the above steps can also be separated or be combined, depending upon the embodiment. In some cases, the steps can also be changed in order without limiting the scope of the invention claimed herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. It is also understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

FIG. **4A** is a simplified diagram of a computing system **430** according to an embodiment of the present invention. This diagram is merely an example which should not unduly limit the scope of the claims herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. Like reference numerals are used in this FIG., as the previous FIG. for cross-referencing purposes only. As shown, the computing system **430** carries out certain functionality that is integrated into the method above as well as others. The computing system includes an accounting module **429**, which carries out certain accounting functions. The accounting module interfaces with mass memory storage **431**, a microprocessing device **433**, and a network interface device **435**, which couples to local and/or wide area networks. The module oversees an invoicing step **417** and transfer step **427**, as shown. Here, the accounting module is a task master for the service based method for using the content delivery network and/or global traffic management network.

Before discussing the accounting module in detail, we begin an overall method at start, step **401**. The method connects (step **403**) a client to a server location through a world wide network of computers. The world wide network of computers can include an internet, the Internet, and others. The connection occurs via a common protocol such as TCP/IP. The client location is coupled to a server, which is for a specific user. The user can be any web site or the like that distributes content over the network. As merely an example, the user can be a portal such as Yahoo! Inc. Alternatively, the user can be an electronic commerce site such as Amazon.com and others. Further, the user can be a health site. Information sites include the U.S. Patent Office web site, educational sites, financial sites, adult entertainment sites, service sites, business to business commerce sites, etc. There are many other types of users that desire to have content distributed in an efficient manner.

In a specific embodiment, the user registers its site on the server, which is coupled to a content distribution server coupled to a global traffic management server. The user registers to select (step **407**) a service from the server. The service can be either a traffic management service (step **414**) or a traffic management service and content distribution service (step **411**). Depending upon the embodiment, the user can select either one and does not need to purchase the capital equipment required for either service. Here, the user merely registers for the service and pays a service fee. The service fee can be based upon a periodic time frequency or other parameter, such as performance, etc. Additionally, the user enters information such as the user's domain name, physical address, contact name, billing and invoicing instructions, and the like. Once the service has been requested, the user performs some of the steps noted herein to use the service.

Next, the method processes (step **423**) the user's request and allows the user to use the content distribution network and/or global traffic management network, where the user's web pages are archives and distributed through the content distribution network in the manner indicated herein. The user's web site should become more efficient from the use of such networks. Once a periodic time frequency or other frequency has lapsed (step **419**), the method goes to an invoicing step, step **417**. Here, the method accesses the accounting module, which can retrieve registration informa-

tion about the user, service terms, invoices, accounts receivables, and other information, but is not limited to this information. The accounting module determines the service terms for the user, which has already registered. Once the service terms have been uncovered from memory, the module determines the way the user would like its invoice. The accounting module directs an invoicing step, which sends (step **427**) an invoice to the user. Alternatively, the process continues until the periodic time frequency for the designated service lapses via line **422**. The invoice can be sent via U.S. mail, electronic mail, or the like. The method stops, step **425**. Alternatively, the invoicing step can deduct monetary consideration through an electronic card, e.g., debit card, credit card. To finalize the transaction, an electronic mail message can be sent to the user, which is logged in memory of the accounting module.

This sequence of steps is merely illustrative. The steps can be performed using computer software or hardware or a combination of hardware and software. Any of the above steps can also be separated or be combined, depending upon the embodiment. In some cases, the steps can also be changed in order without limiting the scope of the invention claimed herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. It is also understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

EXAMPLE

To prove the principle and operation of the present invention, we have provided examples of a user's experience using the present invention. These examples are merely for illustration and should not unduly limit the scope of the claims herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. For easy reading, we have provided a description for a user's experience of a content delivery network and a user's experience of a global traffic management service, which is coupled to such content delivery network.

Content Delivery Network

1. Overview

In a specific embodiment, the invention provides a content distribution network. The following description contains information on how to use a graphical user interface to monitor activity, control cache, and perform checks. In some embodiments, the invention also provides a way for customer feedback to improve the service.

The present network is substantially always available in preferred embodiments. The network includes a Network Operations Center (NOC), which is dedicated to maintaining the highest possible network availability and performance. In most cases, the network is supported and staffed by specially trained service engineers, the 24-hour, 7 day NOC provides consultation, troubleshooting, and solutions for every issue. The staff can be reached through telephone, email, fax, or online. The staff generally connects you to engineers and solutions, not to answering machines.

In a specific embodiment, the network service can be used as long as the user has certain desires. For example, the user has content that needs to be delivered to end-users. This content can be delivered through HTTP, HTTPS, Streaming Media, or FTP, and the like. The server is for hosting the content on the Internet. For standard web content, we implemented a caching system to distribute web content from an origin server to a cache server that is close to a user. This means an origin server needs to exist that contains a master copy of the content. If the user has an existing Web site, the existing Web site will be the origin site.

In one embodiment, the present network is comprised of clusters of servers at points of presence located on many different backbone networks around the world. The servers provide global traffic management and distribution services for content of many kinds, including support for HTTP, HTTPS, FTP, and multiple varieties of streaming media.

In a specific embodiment, the present network includes one or more services. Here, the network may offer services, including:

1. Global Traffic Management—Provides global load balancing across multiple origin sites, along with intelligent failover and other advanced capabilities such as persistence and static mapping.
2. Content Delivery Network (CDN)—Supports content distribution and delivery for HTTP, HTTPS and FTP.
3. Streaming—Supports distribution and delivery of streaming media in many formats, such as Real Media, Windows Media, QuickTime and others.

The present CDN service has some advantages. For example, the CDN service helps increase the performance of any conventional Web site or other Internet services. It also helps reduce latency problems and packet loss, and it provides for content synchronization and replication. The network also reduces latency problems and packet loss. Latency problems result when the user's request travels beyond a certain distance or makes a number of network hops. When users request content from the web or FTP sites, the requests are routed through the Internet to the server. If, as is true for many companies, the servers are located at only one site or a small number of sites, they will not be close to most of the users. Therefore, the users' request for content might traverse many networks to communicate with the desired servers.

Latency problems are often aggravated by packet loss. Packet loss, common on the Internet, tends to worsen at "peering points," locations where different networks connect. One way to reduce packet loss and latency is to install content servers closer to users and ensure that when a user requests data, the request is routed to the closest available server. The present network has deployed web caches, streaming, and FTP servers throughout the Internet, on many networks close to end users. In addition, the network uses a Global Traffic Manager that routes traffic to the closest, most available and least loaded server.

The network often synchronizes the content on the customer's origin site with the Web cache servers on the network. When new content is placed on an origin site and when users make requests for that content, it is automatically replicated to Web cache servers in the network. When new content is published on the origin site with a new name, it is generally immediately available from all caches in the present network. For example, the network user might add an object to the site where a similar object exists:

Add "www.customer.com/images/picture2.jpg" to the same site as "www.customer.com/images/picturejpg."

When a request for "picture2.jpg" arrives at a cache the first time, the cache in the network determines that it does not have a copy of "picture2.jpg", and the cache will request a copy from the origin site. To keep in synchronization with the origin site, the caches periodically check the content they have cached against the copy of the content in the origin site. For Web content, this is accomplished by periodically performing an "If-modified-since" request back to the origin

site to see if the content has changed. This causes content changed on the origin site to be refreshed on the caches at a predefined interval. This interval can be configured depending upon ones needs.

The periodic checking is a common feature of caches but if a piece of content is updated, the old content may be invalidated and the new content published to all the caches in the network. The present CDN service makes this purging possible with a cache control utility that allows you to invalidate a single object, a content directory, or an entire site contained in the caches. In a specific embodiment, cache control is available as part of the service—a service provided to all customers. The present service method provides a comprehensive set of monitoring and administration capabilities for management of the web site.

In a specific embodiment, the present service method runs on a secure server on the Internet and can be accessed only through a web browser that supports secure connections (SSL). A username and password are often assigned to a user or customer when signed up for the service.

One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. The above example is merely an illustration, which should not unduly limit the scope of the claims herein. It is also understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

2. Procedures

We now describe the procedures that can perform to set up the present CDN service and to monitor the performance of the Web site:

A. Implementing the CDN;

B. Invalidating content by controlling cache;

C. Monitoring activity; and

D. Performing tests.

Details of each of these procedures are provided below.

A. Implementing the CDN

To implement the CDN, the customer only need to make minor changes to the web pages in order to direct user requests to the present Web caches instead of to the origin site. In a specific embodiment, the method is as simple as changing the pointers in the HTML. When a cache gets a request for content, it will return the requested object if it exists in the cache. If the content does not exist, it will retrieve the content from the origin site and return it to the user, as well as cache the content so that subsequent requests for that object are instantly available.

To modify the site, the customer can either: (1) changing the URL; or (2) set up virtual hosting. In a specific embodiment, the site can be modified for redirecting a user requests by changing the URL in the HTML. The following example, a request for a picture, shows the original html and the revised html.

Original homepage

The original homepage contains the following URL:

http://www.customer.com/page.html

The URL contains the following HTML:

<html><body>

Here is a picture:

<img src="images/picture.jpg">

</body></html>

Revised homepage

The "img src" tag has been revised:

<html><body>

Here is a picture:

<img src="http://customer.speedera.net/www.customer.com/images/picturejpg">

</body></html>

With the original configuration, a user's browser requests the picture from the customer.com Web servers:

page.html from www.customer.com

images/picture.jpg from www.customer.com

With the revised configuration, a user's browser requests the picture from the customer.speedera.net Web servers:

page.html from www.customer.com

www.customer.com/images/picture.jpg from customer.speedera.net

Note: If the cache does not hold the requested object in memory or on disk, it makes a request to the origin site and caches it.

In an alternative embodiment, the method can set up virtual hosting so that the user's request for content is directed to the present CDN instead of to the origin site. Here, the customer can change the DNS setup to cause the domain name to resolve to the present network cache servers instead of to the original Web server. The domain name may be changed, for example, change the domain name from www.customer.com to wwx.customer.com. The present caches in the network can be configured in a way such that when they get a request for www.customer.com content they have not cached, they can make a request to the wwx.customer.com origin site to get the content. Here, the URLs in the Web pages may not need to be changed.

B. Invalidating Content by Controlling Cache

To invalidate the content contained in the caches, do the following:

1. Access the user interface at: https://speedeye.speedera.com
2. Find the Cache Control page (see FIG. 5A) in the Content Delivery section of the interface.
3. Enter the URL in the text field.
4. Click Submit.

For example, if an image:

www.customer.com/images/picture.jpg and the user changed the image without changing the name and the user wanted the change to be immediately reflected in all caches in the network, the user could use the present service to invalidate the old content, as follows:

Enter "http://www.customer.com/images/picture.jpg" to invalidate the individual picture, or "http://www.customer.com/images/" to invalidate all content in the images directory, or "http://www.customer.com" to invalidate all content in the domain.

Note: Invalidating any of the above causes the change to "picture.jpg" to immediately be reflected in all the caches.

C. Monitoring Activity

In a specific embodiment, the present method allows the user to monitor the operation of the Content Delivery Network service. The present method shows how much content is being delivered and where it is being delivered. The start section of the user interface contains a table that shows the present domains and associated origin domains your account is set up to use, as shown in FIG. 5B.

In a specific embodiment, the method includes monitoring recent activity, as shown in FIG. 5C. Here, the user can view the current and last 24 hours of content delivery traffic for a given domain:

1) Access the user interface at:
https://speedeye.speedera.com

2) Find the Recent Activity page in the Content Delivery section of the interface.

As shown, the has more than one graphs. The first shows the amount of traffic served by the content delivery network for that domain over the last 24 hours. The current traffic is shown on the far right. A dotted vertical line separates data from yesterday on the left and data from today on the right. A second graph on the same page (see FIG. 4) shows the number of hits per second over the last 24 hours. The total number of hits over the last 24-hour period is shown in the title bar of the graph.

In an alternative embodiment, the method includes monitoring activity by location Here, the user views the last 24 hours of content delivery traffic by location for a given domain:

1. Access the user interface at:
https://speedeye.speedera.com

2. Find the By Location page in the Content Delivery section of the user interface.

A world map appears (see FIG. 5D) that shows all the locations that served traffic for the domain.

Below the world map is a bar graph (see FIG. 5E) that shows the amount of traffic served from each individual location over the last 24 hours for a given domain name. This graph is useful for many purposes, such as for determining the optimal location for a second origin site—typically, at the location serving the most traffic, where there is not currently an origin site and when that location is on a different network than the existing origin site.

D. Performing Tests

According to the present invention, selected tests can be performed to check performance, as follows:

1) Access the user interface at:
https://speedeye.spedera.com

2) Locate the Tests section.

3) Select the test you want to perform.

A "page check" test can be performed. This test allows the user to check the performance of a Web page from multiple locations. To use the page check program, do the following:

1) In the text field, enter the URL to test.

2) Select the locations from which the user wants to check the page.

3) Click Check.

At that point, servers at the location(s) selected will be contacted to hit the Web page associated with the URL entered and time how long it takes to download the page and all its components. When the servers have completed downloading the page, the results are shown in the form of tables and graphs. The first table (see FIG. 5F) is the overall performance table. It appears at the top of the results.

In this example, the page took an average of 500 milliseconds (half a second) to download from the first three locations (rows) and 1317 milliseconds (1.3 seconds) from the last location. A server name, physical location, and network location identify each location. For example, the last location in FIG. 5G is labeled as "server-4/sterling/exodus." This label identifies a server on the Exodus network located in Sterling, Va., USA.

After the overall timetable, details for each location are presented in individual tables. FIG. 5H shows a table containing the details for the location "server-14, dc, cw, a server located on the Cable & Wireless Network in Washington D.C., USA. The IP address of the actual server is shown in the heading of the table so you can perform additional tests, if needed, (traceroute and so on) on the actual server performing the test.

The Location table in FIG. 5H shows data for the www.speedera.com Web site. The graph shows the performance for downloading specific components of the page. This table shows that the majority of the time spent in the download was spent downloading the home page itself. The remainder of the content (all the gifs on the subsequent lines) has been cached and is delivered from the closest and least loaded available server within the CDN, in a fraction of the time. These cached items have a domain name of www.speedera.net.

In a specific embodiment, the colors in the graph show the different components of the download including the DNS lookup time, connect time, and so on. The first time a page is checked, the DNS times will likely be very high. This high reading results from the way DNS works in the Internet. If a domain name is not accessed within a specific amount of time (the timeout period), the information will expire out of the DNS caches. The first request will again need to walk through the Internet's hierarchical system of DNS servers to determine which one is authoritative for a given domain name.

To get even more accurate results, a page can be hit twice, where the results from the second hit are used. This will give a more accurate representation of what the performance is like when the page is being hit on a regular basis. The graph is followed by the actual raw data that makes up the graph. Each row displays the following elements:

URL. The URL component downloaded

IP Address. The IP address of the server contacted to get the data

ERR. The error code (where 0 is no error)

HRC. The HTTP response code (where 200 is OK)

LEN. The length of the data downloaded

CHK. A checksum of the data

STT. The timing in milliseconds for the start time

DRT. DNS response time in milliseconds

COT. Connection Time—Syn/SynAck/Ack Time

DST. Data start time when first packet is downloaded

FNT. Final time when download is complete

END. The total millisecond timings for portions of the connection

Global Traffic Manager

The present invention provides a global traffic manager. The global traffic manager is coupled to the content delivery network. The following provides a description of the global traffic manager. The description is merely an illustration, which should not unduly limit the claims herein. One of ordinary skill would recognize many other variations, alternatives, and modifications.

1. Procedures

To use the Global Traffic Management service, the following will be used:

A. Domain name representing a service.

The domain name can be delegated for which the users are authoritative so that the present servers are contacted to resolve the domain name to an IP address, or addresses. Alternatively, we can create a domain name for you. That name will end with speedera.net, such as customer.speedera.net.

B. More that one IP address associated with that service.

Obtaining more that one IP address for a given service provides the following benefits from the Global Traffic Management service:

Provides better service for clusters of servers on multiple networks. If a location within a cluster fails, or the network associated with that location fails, the system can route traffic to another available network because there is more than one IP address. The system also provides better performance by sending user requests to the closest cluster of servers. These routing options are not available if a local load balancer is used to manage the cluster, since a local load balancer requires that each cluster of servers use a single IP address.

Provides better service for clusters of servers on a single network. If each computer has a different IP address, the Global Traffic Management service can be used to load-balance between individual computers.

Reduces latency for a single cluster of servers that is attached to multiple network feeds. In this configuration, the Global Traffic Management can route around network failures by testing each of the network connections and by routing user requests to the closest working connection.

In a specific embodiment, the present network is comprised of clusters of servers at points of presence located on many different backbone networks around the world. The servers provide global traffic management and distribution services for content of many kinds, including support for HTTP, HTTPS, FTP, and multiple varieties of streaming media. As previously noted, the services include: Global Traffic Management—Provides global load balancing across multiple origin sites, along with intelligent failover and other advanced capabilities such as persistence and static mapping; Content Delivery Network (CDN)—Supports content distribution and delivery for HTTP, HTTPS and FTP; and Streaming—Supports distribution and delivery of streaming media in many formats, such as Real Media, Windows Media, QuickTime and others.

The present Global Traffic Management service routes user requests to the closest available and least-loaded server. The service also tests the servers it manages for service performance and availability, using actual application-level sessions. When a service test fails, the system reroutes the traffic to other available servers. The Global Traffic Management service is based on Domain Name Service (DNS). The Internet uses the DNS to allow users to identify a service with which they want to connect. For example, www.speedera.com identifies the Web service (www) from speedera.com.

When users request a service on the Internet, they request it by its DNS name. DNS names were created to make it easier for users to identify computers and services on the Internet. However, computers on the Internet do not communicate with each other by their DNS names Therefore, when a user enters a domain name, domain name servers on the Internet are contacted to determine the IP addresses associated with that name.

The Network includes specialized domain name servers that use advanced mechanisms to determine the IP addresses associated with a given domain name and service. These servers work seamlessly with the Internet DNS system. To determine the best IP address, or addresses, to return when a user requests a service on the Internet, the DNS system does the following:

1. Uses IP addresses to monitor the performance of a service on individual computers or clusters of computers
2. Determines latency and load metrics between users and servers on the Internet
3. Performs tests on the Internet to determine the quality of service a user would receive when connecting to a specific computer or cluster of computers

Procedures

This section describes the procedures you can perform to implement and then monitor the performance of the Global Traffic Management service. To implement the Global Traffic Management service, the customer or user does the following:

1. Sign up for the service.
2. Contact the server location and provide the following information: The domain name of the service you want the system to manage; The IP addresses associated with that service; A description of the service and how it should be tested for performance and availability; The interval after which tests should be performed; What the service check should look for, such as specific information in a returned Web page. Whether the user would like traffic weighted so that more traffic is sent to one IP address over another.

In addition to the normal routing around failures to the closest server, the system can also be set up for security purposes. The system can contain hidden IP addresses that are only given out in the case of failure of other IP addresses. The user might want to use this feature to prevent a denial of service attack. If one IP address is attacked and becomes unavailable, another will then appear and traffic will be routed to it. This can make attacking a Web server more difficult since the IP address is not published until the failure occurs.

In a specific embodiment, the method allows the user to monitor the operation of the Global Traffic Management service for domain names. Preferably, the method outputs information on a Web-based, user-interface that runs on a secure server on the Internet that can be accessed only through a web browser that supports secure connections (SSL). Here, a start section of the user interface contains a table that shows all the domains and associated origin domains your account is set up to use. See FIG. 6A.

In an alternative embodiment, we can also view the last 24 hours of traffic management activity for a given domain:

1) Access the user interface at:
https://speedeye.speedera.com
2) Find the Recent Activity page in the Traffic Management section of the interface.

The main graph in the page shows how traffic was routed over the last 24 hours. A dotted vertical line separates yesterday on the left from today on the right. The lines in the graph show how many times each IP address was given out. See the example in FIG. 6B.

In the example, the present Global Traffic Management system made 198120 traffic routing decisions over a 24-hour period. The lower decision line in the graph represents an IP address for "Delhi, India." The upper decision line represents an IP address for "Santa Clara, Calif.; United States." The Y axis represents the activity levels. The X axis represents the Santa Clara time: N for noon, P for p.m., and A for a.m.

At 6:00 a.m. in Santa Clara, one line dropped to the bottom of the graph and the other spiked upward. This happened because the system routed around a failure at a data center. When the "Delhi" IP address failed its service test, the Global Traffic Management system routed traffic to the "Santa Clara" IP address.

The example also shows that the "Delhi" IP address is more active at night (Santa Clara time), and the "Santa Clara" IP address is more active in the daytime. The difference in activity results from the changes in time zones. When people in India are active, the traffic manager routes their requests to the closest available server with the best

service response time. For users in India, when it is their daylight and their peak time, the best IP address is often the site in Delhi. For users in the U.S., when it is their peak time, the best IP address is the site in Santa Clara.

In still an alternative embodiment, we can view the last 24 hours of traffic management activity by location for a given domain:

1. Access the user interface at: https://speedeye.speedera.com
2. Find the By Location page in the Content Delivery section of the user interface.

Here, a world map and a bar chart appear. They show where the traffic manager routed traffic (geographic and network locations) over the last 24 hours for a given domain name. See the example in FIG. 6C. The bar-chart example shows the number of times each location was chosen to serve traffic over the last 24 hours. In the example, the traffic manager chose the "UUNET/sclara"(Santa Clara, Calif.; United States) location to serve most of the traffic.

In other aspects, the method includes performing tests. Here, the interface also contains a utility that allows the user to check a Web page from multiple locations. If an HTTP service is used, a quick status check can be executed as follows:

1) Access the user interface at: https://speedeye.spedera.com
2) In the text entry field, enter the URL for the page you want to check.
3) Select the locations from which you want to check the page.
4) Press the Check button. This causes servers at the location, or locations, selected to download the Web page associated with the URL you entered in Step 2.

When the servers have completed downloading the page, the page-performance results are shown in the form of tables and graphs. The first table (see FIG. 6D) is the overall performance table. It appears at the top of the results. In this example, the page took an average of 500 milliseconds (half a second) to download from the first three locations (rows) and 1200 milliseconds (1.2 seconds) from the last location.

A server name, physical location, and network location identify each location. For example, the last location in FIG. 6D is labeled as "server-4/sterling/exodus." This label identifies a server on the Exodus network located in Sterling, Va., USA.

After the overall timetable, details for each location are presented in individual tables. FIG. 5 shows a table containing the details for the location "server-14, dc, cw, a server located on the Cable & Wireless Network in Washington D.C., USA. The IP address of the actual server is shown in the heading of the table so you can perform additional tests, if needed, (traceroute and so on) on the actual server performing the test. The Location table in FIG. 6E shows data for the www.speedera.com Web site.

The graph in FIG. 6E shows the performance for downloading specific components of the page. This table shows that the majority of the time spent in the download was spent downloading the home page itself.

The colors in the graph show the different components of the download including the DNS lookup time, connect time, and so on. The first time you check a page, the DNS times will likely be very high. This high reading results from the way DNS works in the Internet. If a domain name is not accessed within a specific amount of time (the timeout period), the information will expire from the DNS caches. The first request will again need to walk through the Internet's hierarchical system of DNS servers to determine which one is authoritative for a given domain name.

To get more accurate results, a page can be hit twice and the results from the second hit can be used. This will give you a more accurate representation of what the performance is like when the page is being hit on a regular basis. In the Location Table, the graph is followed by the actual raw data that makes up the graph. Each row displays the following elements:

URL. The URL component downloaded

IP Address. The IP address of the server contacted to get the data

ERR. The error code (where **0** is no error)

HRC. The HTTP response code (where **200** is OK)

LEN. The length of the data downloaded

CHK. A checksum of the data

STT. The timing in milliseconds for the start time

DRT. DNS response time in milliseconds

COT. Connection Time—Syn/SynAck/Ack Time

DST. Data start time when first packet is downloaded

FNT. Final time when download is complete

END. The total millisecond timings for portions of the connection

In a specific embodiment, the Global Traffic Management system automatically routes around failures to services on the IP addresses it manages. Here, the system can also be: Adding or removing a domain name from the system; Adding or removing IP addresses from the system; and Changing the way a service is monitored.

It is also understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference for all purposes in their entirety.

What is claimed is:

1. A service based system for traffic management and content distribution for a plurality of users over a world wide network of computers, the system comprising:

a global traffic management device coupled to a world wide area network, the global traffic management device being provided to load balance traffic for a customer across multiple customer origin sites;

a content delivery network coupled to the global traffic management device, the content delivery network providing content distribution and delivery of any of: static Web page content or streaming media, for the customer;

a computing device including a computer memory coupled to the global traffic management device; and

an accounting module coupled to the computing device, the accounting module being provided to track a usage of the global traffic management device and the content delivery network for the customer of the global traffic management device and the content delivery network to determine a service fee to the customer of the global traffic management device and the content delivery network for the usage based upon a period time frequency.

2. The system of claim 1 wherein the periodic time frequency is selected from an hour, a week, a month, a quarter, or a year.

3. The system of claim 1 wherein the content distribution network comprises a plurality of content servers, each of the content servers being spatially disposed in a geographical manner.

4. The system of claim 1 wherein the content distribution and delivery is for at least HTTP, HTTPS and FTP.

5. The system of claim 1 wherein the streaming media is selected from Real Media, Windows Media, and QuickTime.

6. The system of claim 1 wherein the global traffic management system comprises a plurality of management agents being distributed throughout the content distribution network.

7. The system of claim 1 wherein the customer of the content distribution network is an Internet commerce company.

8. The system of claim 1 wherein the usage of the content distribution network reduces a latency of web pages of the customer to a user.

9. The system of claim 1 wherein the global traffic management device reduces a latency of web pages of the customer.

10. The system of claim 1 wherein the service fee is a one time fee.

11. The system of claim 1 wherein the global traffic management system provides intelligent failover for origin sites.

12. A service based system for traffic management and content distribution for a plurality of users over a world wide network of computers, the service based system being free from capital expenditure to the plurality of users and being provided on periodic service fees, the system comprising:

a global traffic management device coupled to a world wide area network, the global traffic management device being provided to load balance traffic for a customer across multiple customer origin sites;

a content delivery network coupled to the global traffic management device, the content delivery network providing content distribution and delivery of any of: static Web page content or streaming media, for the customer;

a computing device including a computer memory coupled to the global traffic management device; and

an accounting module coupled to the computing device, the accounting module being provided to track a usage of the global traffic management device and the content delivery network for the customer of the global traffic management device and the content delivery network to determine a service fee to the customer of the global traffic management device and the content delivery network for the usage based upon a period time frequency, the accounting module comprising:

a first code directed to retrieving invoicing information from the computer memory; and

a second code directed to determining a service fee based upon the invoicing information from the computer.

13. The system of claim 12 wherein the invoicing information comprises an address and a service fee.

14. The system of claim 12 wherein the service fee is based upon a periodic frequency.

15. The system of claim 12 wherein the content distribution and delivery is for at least HTTP, HTTPS and FTP.

16. The system of claim 12 wherein the customer is free from paying a capital fee toward the usage of the content delivery network and the global traffic management device.

17. The system of claim 12 wherein the customer of the content distribution network is an Internet commerce company.

18. The system of claim 12 wherein the Internet commerce company is a book store, a business to business site, a portal, a health site, and a finance site.

19. The system of claim 12 wherein the usage of the content distribution network reduces a latency of web pages of the customer to a user.

20. The system of claim 12 wherein the customer of the content distribution network is an Internet commerce company.

21. The system of claim 12 wherein the periodic time frequency is selected from an hour, a week, a month, a quarter, or a year.

22. The system of claim 12 wherein the global traffic management system provides intelligent failover for origin sites.

* * * * *